CS534 - Machine Learning

Homework Assignment 0

Author: Vy Bui
OSUID: 934370552
Email: buivy@oregonstate.edu

The School of Electrical Engineering and Computer Science
Oregon State University

Linear Algebra

(1a) The product is not defined because the neighboring dimensions do not match.

(1b)

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1\times1+2\times0+3\times1 & 1\times1+2\times1+3\times0 & 1\times0+2\times1+3\times1 \\ 4\times1+5\times06\times1 & 4\times1+5\times1+6\times0 & 4\times0+5\times1+6\times1 \\ 7\times1+8\times0+9\times1 & 7\times1+8\times1+9\times0 & 7\times0+8\times1+9\times1 \end{bmatrix}$$

$$= \begin{bmatrix} 4 & 3 & 5 \\ 10 & 9 & 11 \\ 16 & 15 & 17 \end{bmatrix}$$

(1c)

$$\begin{bmatrix} 1 & 2 & 1 & 2 \\ 4 & 1 & -1 & -4 \end{bmatrix} \begin{bmatrix} 0 & 3 \\ 1 & -1 \\ 2 & 1 \\ 5 & 2 \end{bmatrix} = \begin{bmatrix} 1\times0+2\times1+1\times2+2\times5 & 1\times3+2\times(-1)+1\times1+2\times2 \\ 4\times0+1\times1-1\times2-4\times5 & 4\times3+1\times(-1)-1\times1-4\times2 \end{bmatrix}$$

$$= \begin{bmatrix} 14 & 6 \\ -21 & 2 \end{bmatrix}$$

(2a)

$$A = \begin{bmatrix} 1 & 1 & -1 & -1 \\ 2 & 5 & -7 & -5 \\ 2 & -1 & 1 & 3 \\ 5 & 2 & -4 & -2 \end{bmatrix}$$

$$b = \begin{bmatrix} 1 \\ -2 \\ 4 \\ 6 \end{bmatrix}$$

(2b) Because A is a square matrix and $det(A) \neq 0$, A is nonsingular, therefore invertible. And because A is nonsingular,

$$A^{-1} = \begin{bmatrix} 0.5 & -0.167 & 0 & 0.167 \\ 2 & 0.167 & 0.5 & -0.667 \\ 1.75 & -0.25 & 0 & -0.25 \\ -0.25 & 0.25 & 0.5 & -0.25 \end{bmatrix}$$

$$x = A^{-1}b = \begin{bmatrix} 0.5 & -0.167 & 0 & 0.167 \\ 2 & 0.167 & 0.5 & -0.667 \\ 1.75 & -0.25 & 0 & -0.25 \\ -0.25 & 0.25 & 0.5 & -0.25 \end{bmatrix} \begin{bmatrix} 1 \\ -2 \\ 4 \\ 6 \end{bmatrix} = \begin{bmatrix} 1.83 \\ -0.33 \\ 0.75 \\ -0.25 \end{bmatrix}$$

Vector Calculus

(1a)
Let $g(x) = 1 + e^{-x}$ and $h(g) = g^{-1}$, then we have $f(x) = h(g(x))$.
$f'(x) = h'(g)g'(x) = (-g^{-2})g'(x) = -\frac{1}{(1+e^{-x})^2}(-e^{-x}) = \frac{1}{1+e^{-x}}(1 - \frac{1}{1+e^{-x}}) = \sigma(x)(1 - \sigma(x))$

(1b)
$f'(x) = e^{-\frac{(x-\mu)^2}{2\sigma^2}}(-\frac{2(x-\mu)}{2\sigma^2}) = \frac{\mu - x}{\sigma^2 e^{\frac{(x-\mu)^2}{2\sigma^2}}}$

(2a)

We have

$$z = x^T x = \begin{bmatrix} x_1 & x_2 & ... & x_D \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix} = x_1^2 + x_2^2 + ... + x_D^2$$

Applying chain rule results in
$\nabla_x f = \frac{\partial f}{\partial x} = \frac{\partial f}{\partial z}\frac{\partial z}{\partial x} = \frac{1}{1+z}$
$\begin{bmatrix} \frac{\partial z(x)}{\partial x_1} & \frac{\partial z(x)}{\partial x_x} & ... & \frac{\partial z(x)}{\partial x_D} \end{bmatrix} = \frac{1}{1+x_1^2+x_2^2+...+x_D^2} \begin{bmatrix} 2x_1 & 2x_2 & ... & 2x_D \end{bmatrix}$
Because $z = x^T x$ is a scalar, $f(z(\mathbf{x}))$ is also a scalar. In other words, $\mathbf{f} : R^D \to R^1$ maps vector $\mathbf{x}$ to a scalar. Therefore, the gradient of f with respect to $\mathbf{x}$ is a $1 \times D$ matrix, which can also be seen as the result of the gradient above.

(2b) According to chain rule, we have
$\nabla_x f = \frac{\partial f}{\partial x} = \frac{\partial f}{\partial z}\frac{\partial z}{\partial x}$ (1)
We also have $\frac{\partial f}{\partial z} = -\frac{1}{2e^{z/2}}$ (2)
Because S is a symmetric matrix, S^{-1} is also a symmetric matrix, which enables us to apply formula (85) in the matrix cookbook as following
$\frac{\partial z}{\partial x} = \frac{\partial}{\partial x}(\mathbf{x} - \mu)^T S^{-1}(\mathbf{x} - \mu) = 2S^{-1}(\mathbf{x} - \mu)$ (3)
$(1), (2), (3) \to \nabla_x f = -\frac{S^{-1}(\mathbf{x}-\mu)}{e^{z/2}}$

Probability

(1a)
Let the sample space $\Omega = \{Fair, Unfair\}$, target space T $= \{T, F\}$ denoting whether the picked coin is fair or not, and random variable $X : \Omega \to$ T.

Assume that the difference between two coins is so small and cannot be recognized, then $P(X = T) = P(Fair) = 0.5$.

(1b)
Let us define a sample space $\Omega = \{HH, HT, TH, TT\}$ denoting the outcome of two tosses. The event we are interested in is whether the first toss is head or not. Let us define a random variable Y that maps Ω to T $= \{H, T\}$, which denotes the result of the first toss.

$P(Y = H) = P(Y = H|X = F) + P(Y = H|X = T)$
$= P_F(H)P(X = T) + P_U(H)P(X = F) = 0.5 \times 0.5 + 0.1 \times 0.5 = 0.3$

(1c)
Applying Bayes's theorem, we have $P(X = T, HH) = \frac{P(HH|X=T)P(X=T)}{P(HH|X=T)P(X=T)+P(HH|X=F)P(X=F)} = \frac{(0.5\times0.5)\times0.5}{(0.5\times0.5)\times0.5+0.1\times0.1\times0.5} \approxeq 0.962$

(2a)
By convention, $p(x) = 1$ and $p(y) = 1$ for discrete random variables with a finite number of events.

(2b)
$p(x|Y = y_1) = 0.01 + 0.02 + 0.03 + 0.1 + 0.1 = 0.26$

$p(y|X = x_3) = 0.03 + 0.05 + 0.03 = 0.11$

(3a)
The likelihood function can be written as $\Pi_{i=1}^{n} f(x_i; 0, \theta) = \Pi_{i=1}^{n} \frac{1}{\theta - 0} = \frac{1}{\theta^n}$

(3a)
Because the likelihood function is a decreasing function of θ, the estimate is smallest when θ is largest, that is the MLE of θ, $\hat{\theta} = max(x_1, x_2, ..., x_n)$.

(4a)
The cost of filtering the mail out is cost of incorrectly predict it as spam; $Cost = 10 \times (1 - P) + 0 \times P = 10 \times 0.2 = 2$.

(4b)
The cost of not filtering out the email is cost $= 0 \times 0.2 + 1 * 0.8 = 0.8$.

Because it is cheaper not to filter out the email, choose to label the mail as non-spam.

(4c)
Applying the threshold will make the cost of filtering out $= 10 \times (1 - \theta)$, and the cost of not filtering out $= 1 \times \theta$. The threshold should balance of these two costs, that is

$\theta = 10 \times (1 - \theta) \to \theta = 10/11$.

By doing so, we embed the difference in cost of these two decisions in θ. The result aligns with

our observation from 4a and 4b when false negative cost is cheaper than that of false positive. Increasing threshold helps decrease false positive, which in turn reduces false positive cost.